



# **Projet final**

## **Solr**

Présenté par

Chengwanli YANG

Master 2

Indexation sémantique et recherche d'information

Sorbonne Université – Faculté des Lettres  
Novembre 2021

## 1. Introduction

Aujourd'hui au fur et à mesure du développement de l'informatisation, les informations massives offrent une grande commodité pour répondre à nos besoins. Cependant, la diversité et la variabilité des informations rendent difficiles pour que les gens cherchent rapidement et précisément les informations surtout sur l'internet. Par contre, le moteur de recherche pourrait trouver efficacement les ressources de façon automatique. Solr qui est une plateforme logicielle de moteur de recherche se fonde sur la bibliothèque Lucene, développée par l'Apache Software Foundation<sup>1</sup>. Elle pourrait communiquer avec l'utilisateur à l'aide d'une interface « solr admin » en réalisant les 4 tâches principales<sup>2</sup> :

1. L'indexation : convertir les documents dans un format lisible par la machine.
2. L'interrogation : comprendre les termes d'une requête posée par l'utilisateur. Ces termes peuvent être des images ou des mots-clés.
3. Le mappage : Mapper la requête de l'utilisateur sur les documents stockés dans la base de données pour trouver le résultat approprié.
4. Le classement : dès que le moteur recherche les documents indexés, il classe les sorties selon leur pertinence.

L'objectif de notre travail est de créer un moteur de recherche préliminaire en utilisant Solr. Nous allons d'abord présenter le corpus choisi sur lequel le moteur de recherche sera basé. Ensuite, nous parlerons des configurations du moteur de recherche, le schéma créé, des facettes générées et des modifications de Velocity pour visualiser le contenu des résultats. Finalement, nous donnerons des exemples de notre moteur de recherche.

## 2. Sélection du corpus

Nous avons trouvé le corpus sur le site *data.gouv.fr*<sup>3</sup>, c'est une plateforme ouverte de diffusion des données publiques françaises sous l'autorité du Premier ministre. Le corpus est un jeu de donnée « Enquête sur les ressources électroniques des établissements de l'enseignement supérieur et de la recherche » fourni par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, en novembre 2020.

---

<sup>1</sup> <https://www.apache.org>

<sup>2</sup> [https://en.wikipedia.org/wiki/Apache\\_Solr](https://en.wikipedia.org/wiki/Apache_Solr)

<sup>3</sup> <https://www.data.gouv.fr/fr/>

Il présente les résultats de l'enquête sur les ressources électroniques des établissements de l'enseignement supérieur et de la recherche. Cette enquête permet de recenser les moyens (en €) consacrés par le système documentaire de l'enseignement supérieur et de la recherche à l'acquisition de ressources documentaires électroniques de 2015 à 2018. Le recensement des dépenses produit par l'enquête sur les ressources électroniques couvre plus de 90 % des dépenses en ressources électroniques en France<sup>4</sup>. Le corpus téléchargé est un fichier au format csv qui contient 22349 documents et 21 attributs (millésime, établissement, catégorie, ressource, valeur, académie, département, région, wikidata, etc.). Nous l'expliquons avec un exemple : Sorbonne Université (établissement) est un établissement public à caractère scientifique, culturel et professionnel (catégorie juridique) qui appartient à l'académie parisienne (académie), à Paris (département) et en Île-de-France (région). Cette université (catégorie) a dépensé 4475.48 euros (valeur) en 2018 (millésime) pour ACM Digital Library (ressource). Chaque document est représenté en tabulation dans notre corpus, évidemment, il y a aussi d'autres attributs, comme le code de département, le code d'académie, l'id de paysage, etc.

### **3. Annotation du corpus**

Une fois le corpus choisi, nous pouvons l'annoter pour le prétraitement. Étant donné que notre corpus est pur, il a été nettoyé par le producteur. Sur lequel, il n'y a pas des textes longs ou des données massives, toutes les informations sont claires. Par conséquent, nous pouvons le traiter sans difficulté.

### **4. Création du schéma**

Avant d'indexer le corpus, il est nécessaire de créer un schéma où nous définissons les types des documents. Mais d'abord, nous devons créer notre collection « ResElec ».

La figure 1 présente les types des champs que nous avons définis. Il existe 5 types :

1. String : « academie », « categorie », « code\_ academie », « code\_ departement », « code\_ region », « departement », « id », « grid », « id\_payage », « idref », « region », « ressource », « ror », « siret », « uai », « wikidata ».

---

<sup>4</sup> <https://www.data.gouv.fr/fr/datasets/enquete-sur-les-ressources-electroniques-des-etablissements-de-lenseignement-superieur-et-de-la-recherche/>

Ces champs contiennent aux données courtes qui peuvent être considérées comme les chaînes caractères.

```
</fieldType>
<field name="_nest_path_" type="_nest_path_" />
<field name="_root_" type="string" docValues="false" indexed="true" stored="false"/>
<field name="_text_" type="text_general" multiValued="true" indexed="true" stored="false"/>
<field name="_version_" type="plong" indexed="false" stored="false"/>
<field name="academie" type="string" indexed="true" stored="true"/>
<field name="categorie" type="string" indexed="true" stored="true"/>
<field name="categorie_juridique" type="text_fr" indexed="true" stored="true"/>
<field name="code_academie" type="string" indexed="true" stored="true"/>
<field name="code_departement" type="string" indexed="true" stored="true"/>
<field name="code_region" type="string" indexed="true" stored="true"/>
<field name="departement" type="string" indexed="true" stored="true"/>
<field name="Epst" type="boolean" indexed="true" stored="true"/>
<field name="etablissement" type="text_fr" indexed="true" stored="true"/>
<field name="grid" type="string" indexed="true" stored="true"/>
<field name="id" type="string" multiValued="false" indexed="true" required="true" stored="true"/>
<field name="id_payage" type="string" indexed="true" stored="true"/>
<field name="idref" type="string" indexed="true" stored="true"/>
<field name="millesime" type="plong" indexed="true" stored="true"/>
<field name="negociee" type="boolean" indexed="true" stored="true"/>
<field name="region" type="string" indexed="true" stored="true"/>
<field name="ressource" type="string" indexed="true" stored="true"/>
<field name="ror" type="string" indexed="true" stored="true"/>
<field name="siret" type="string" indexed="true" stored="true"/>
<field name="uai" type="string" indexed="true" stored="true"/>
<field name="valeur" type="pfloat" indexed="true" stored="true"/>
<field name="wikidata" type="string" indexed="true" stored="true"/>
```

figure 1 Types des champs

2. Text\_fr : « categorie\_juridique », « etablissement »

Ce sont des données importantes. Nous avons besoin de les tokeniser pour extraire les mots-clés. Lorsque l'utilisateur exploite ces mots-clés à chercher, le moteur de recherche pourrait renvoyer les résultats pertinents. Puisqu'ils sont en français, nous avons choisi le type « text\_fr » au lieu de « text\_general ». Par exemple, quand nous cherchons le token « Paris », cela trouve les établissements dont le nom comporte « Paris », « Université Sorbonne Nouvelle - Paris 3 », « Université Paris 8 - Vincennes - Saint-Denis », « Institut supérieur de mécanique de Paris (Supméca) », etc.

3. Booléen : « negociee » et « Epst »

Ces 2 champs ne disposent que de 2 valeurs : « true » ou « false ». C'est-à-dire qu'est-ce que la dépense a été négociée ? Oui ou non.

4. Float : « valeur »

Le champ « valeur », autrement dit la dépense est le chiffre avec le virgule, donc c'est le type float.

5. Long : « millesime »

Les années 2015, 2016, 2017, 2018 sont dans le champ « millesime », donc nous avons choisi le type long.

Tous les champs peuvent être indexés, sauvegardés.

### 5. Indexation du corpus

Une fois avoir sauvegardé le schéma que nous avons créé, nous pouvons indexer notre corpus. D’abord, nous avons mis le corpus dans le répertoire « `exampledocs` ». Ensuite, nous avons exécuté la commande ci-dessous sur le terminal.

« `java -Dtype=text/csv -Dc=ResElec -jar post.jar corpus.csv` »

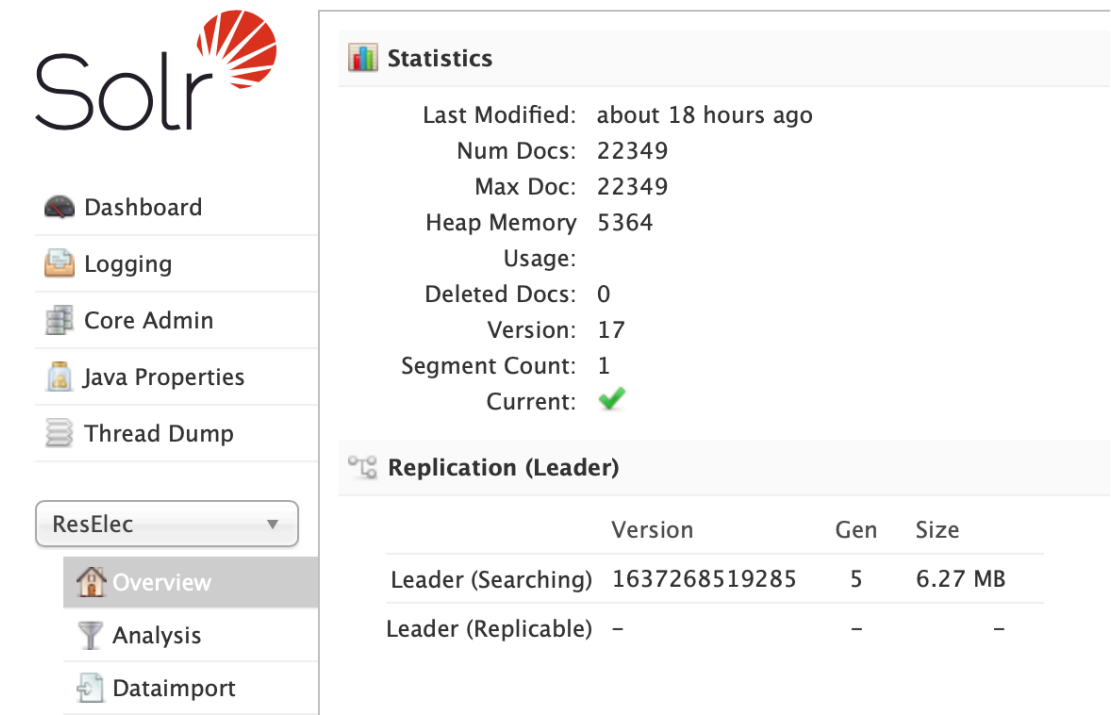


figure 2 Indexation du corpus

```

"millesime":2016,
"categorie_juridique":["Établissement public à caractère scientifique, culturel et professionnel"],
"Epst":false,
"categorie":"Université",
"etablissement":["Université Paris 8 - Vincennes - Saint-Denis"],
"ressource":"4n media",
"negociee":false,
"valeur":2500.0,
"code_departement":"D093",
"departement":"Seine-Saint-Denis",
"code_academie":"A24",
"academie":"Créteil",
"code_region":"R11",
"region":"Île-de-France",
"id_paysage":"Uxr7Z",
"wikidata":"Q1194988",
"uai":"0931827F",
"ror":"04wez5e68",
"siret":"19931827000014",
"id":"a7b59a3b-52db-4825-a414-c6560c21c96f",
"_version_":1716800472081235968,
"grid":"grid.15878.33",
"idref":"26403552"},

```

figure 3 Exemple de document

La figure 2 montre l'indexation du corpus dans la collection « ResElec », les 22349 documents sont bien indexés. Pour vérifier les documents, nous avons fait la requête « \*:\* » dans « Query ».

La figure 3 présente un document indexé. Nous constatons d'abord que tous les attributs ont été bien indexés. De plus, les chiffres sont en bleu, les booléens sont en rouge et les chaînes caractères en vert. En outre, les chaînes caractères entre crochets sont en type text\_fr et tokenisés.

## 6. Création des facettes et mise en évidence

Afin de créer les facettes, il est indispensable de modifier le fichier « solrconfig.xml ». Nous avons créé 3 facettes de champ, 2 facettes d'intervalles et 2 facettes de pivot qui sont sur la figure 4.

```

<!-- Faceting defaults -->
<str name="facet">on</str>
<str name="facet.missing">true</str>
<!-- facettes de champ -->
<str name="facet.field">categorie</str>
<str name="facet.field">academie</str>
<str name="facet.field">ressource</str>
<str name="facet.mincount">300</str>
<!-- facettes pivot -->
<str name="facet.pivot">negociee</str>
<str name="facet.pivot">Epst</str>
<!-- facettes d'intervalles -->
<str name="facet.range">valeur</str>
<float name="f.valeur.facet.range.start">0.0</float>
<float name="f.valeur.facet.range.end">5000000.0</float>
<float name="f.valeur.facet.range.gap">10000.0</float>
<str name="facet.range">millesime</str>
<long name="f.millesime.facet.range.start">2015</long>
<long name="f.millesime.facet.range.end">2018</long>
<long name="f.millesime.facet.range.gap">1</long>

```

figure 4 Facettes créées

Les facettes de champ sont « categorie », « academie », « ressource », et pour éviter la grande quantité de facettes affichée sur l'interface, nous avons décidé « facet.mincount 300 », c'est-à-dire qu'il n'affiche que les facettes qui contiennent plus de 300 documents. Ensuite, les facettes de pivot sont « negociee » et « Epst », ils sont des booléens. Les facettes d'intervalles sont « valeur » en float et « millesime » en long, comme nous avons défini dans le schéma. Les intervalles permettent de chercher facilement les informations de chiffre.

La figure 5 montre la configuration de Highlighting. Nous avons décidé de mettre en évidence sur les attributs « etablissemnt », « academie », « categorie », « categorie\_juridique », « ressource » et « region ». Et le type de Highlighting s'appelle coucou qui est défini dans le fichier « main.css », nous allons en parler à la section 7.

```

<!-- Highlighting defaults -->
<str name="hl">on</str>
<str name="hl.fl">etablissement academie categorie categorie_juridique ressource region</str>
<str name="hl.preserveMulti">true</str>
<str name="hl.encoder">html</str>
<str name="hl.simple.pre">&lt;span class="coucou"&gt;</str>
<str name="hl.simple.post">&lt;/span&gt;</str>

```

figure 5 Mise en évidence

Il faut aussi ajouter des balises dans le fichier « solrconfig.xml », comme la figure 6.

```

<lib dir="${solr.install.dir:../../../../}/contrib/velocity/lib" regex=".*\.jar" />
<lib dir="${solr.install.dir:../../../../}/dist/" regex="solr-velocity-\d.*\.jar" />
<queryResponseWriter name="velocity" class="solr.VelocityResponseWriter" startup="lazy">
  <str name="template.base.dir">${velocity.template.base.dir}</str>
</queryResponseWriter>
<initParams path="/update/**,/query,/select,/tvrh,/elevate,/spell,/browse,update">
  <lst name="defaults">
    <str name="df">etablissement</str>
  </lst>
</initParams>

```

figure 6 Balises ajoutées

## 7. Modification de Velocity

Velocity qui est un exemple d'interface de recherche présente plusieurs fonctionnalités utiles, telles que la recherche, les facettes, le highlighting et la saisie semi-automatique, etc.

Notre répertoire « velocity » est issu de l'exemple « techproducts ». Nous avons adapté le fichier « richtext\_doc.vm » et « main.css ». Dans « richtext\_doc.vm », nous avons ajouté des balises pour que les attributs soient renvoyés comme les résultats. Ce sont les attributs « etablissement », « categorie », « categorie\_juridique », « ressource », « valeur », « academie », « millesime », « departement », « region » et « wikidata ».


Dans le fichier « main.css », nous avons ajouté le paramètre de highlighting « coucou » que nous venons de mentionner à la section 6. Les données qui sont mises en évidence sont en rouge, en gras, avec le carré jaune, en taille de police 150%. Nous avons aussi changé la couleur de nom de champs au paramètre « .facet-field ».

Nous avons également modifié plusieurs fichiers de « velocity ». Par exemple le fichier « header.vm » pour changer le logo, les fichiers « facet\_fields.vm », « facet\_pivot.vm », « facet\_ranges.vm » pour changer les facettes en français, dans l'interface, il y a « facette de champ », « facette d'intervalles » et « facette de pivot » au lieu de « facet fields » « facet pivot » « facet ranges ». Les fichiers « pagination\_top.vm », « pagination\_bottom.vm » pour changer la phrase « xx résultats trouvés en x ms Page 1 de xx » en français. Le fichier « query\_form.vm » pour ajouter une option « Valeur par ordre décroissant ». Maintenant, notre interface est tout en français, plus lisible pour les utilisateurs ne comprenant pas l'anglais.



## 8. Exemple d'utilisation

Maintenant, nous donnerons les exemples de la démonstration.



Crée par Chengwanli YANG  
Type de recherche: **Simple** **Spatial** **Par Groupe**

Recherche:

☐ Valeur par ordre décroissant

**Facette de champ** 22349 résultats trouvés en 5 ms Page 1 de 2235

**catégorie**

- Université (11985)
- Organisme de rech... (1697)
- Grand établissement (1492)
- Bibliothèque (1315)
- Institut ou école... (828)
- Non renseigné (716)
- École de commerce... (633)
- École normale sup... (548)
- Grand établisseme... (541)
- École d'ingénieurs (470)
- Centre hospitalier (344)
- missing (0)

**Établissement:** Université Paris 8 - Vincennes - Saint-Denis  
**Catégorie:** Université  
**Catégorie juridique:** Établissement public à caractère scientifique, culturel et professionnel  
**Ressource:** 4n media  
**Valeur (en €):** 2500.0  
**Académie:** Créteil  
**Millésime:** 2016  
**Département:** Seine-Saint-Denis  
**Région:** Île-de-France  
**Wikidata:** Q1194988

figure 7 Interface par défaut

La figure 7 présente l'interface par défaut. Si nous cherchons le mot « Paris », les résultats sont sur la figure 8. Le mot « Paris » est bien mis en évidence dans les attributs « Établissement » et « Académie », mais pas dans « Département » ! C'est parce que nous n'avons pas défini « département » dans la configuration de highlighting (figure 5) !

Recherche:

☐ Valeur par ordre décroissant

**Facette de champ** 10303 résultats trouvés en 13 ms Page 1 de 1031

**catégorie**

- Université (4076)
- Bibliothèque (1273)
- Organisme de rech... (1178)
- Grand établissement (1143)
- Non renseigné (347)
- missing (0)

**academie**

- Paris (5838)
- Lyon (552)
- Brennes (419)
- Bordeaux (391)
- Non renseigné (347)
- Lille (344)
- Toulouse (306)
- missing (38)

**ressource**

- missing (0)

**Facette d'intervalles**

**valeur**

- 0.0 - 10000.0 (7947)
- 10000.0 - 20000.0 (1037)
- 20000.0 - 30000.0 (434)

**millésime**

**Établissement:** Université Sorbonne Nouvelle - **Paris** 3  
**Catégorie:** Université  
**Catégorie juridique:** Établissement public à caractère scientifique, culturel et professionnel  
**Ressource:** ABC-Clio  
**Valeur (en €):** 1351.0  
**Académie:** **Paris**  
**Millésime:** 2017  
**Département:** Paris  
**Région:** Île-de-France  
**Wikidata:** Q571293

**Établissement:** Université Sorbonne Nouvelle - **Paris** 3  
**Catégorie:** Université  
**Catégorie juridique:** Établissement public à caractère scientifique, culturel et professionnel  
**Ressource:** ABC-Clio  
**Valeur (en €):** 1077.0  
**Académie:** **Paris**  
**Millésime:** 2018  
**Département:** Paris  
**Région:** Île-de-France  
**Wikidata:** Q571293

figure 8 Recherche « Paris »

Si nous cochoons l'option « Valeur par ordre décroissant », le moteur de recherche va chercher d'abord les établissements qui appartiennent à l'académie de Paris, ensuite il affichera les résultats par la valeur ordre décroissant, comme la figure 9.

Recherche:

☒ Valeur par ordre décroissant

**Facette de champ**

**catégorie**

- [Université](#) (4076)
- [Bibliothèque](#) (1273)
- [Organisme de rech...](#) (1178)
- [Grand établissement](#) (1143)
- [Non renseigné](#) (347)
- [missing](#) (0)

**academie**

- [Paris](#) (5838)
- [Lyon](#) (552)
- [Rennes](#) (419)
- [Bordeaux](#) (391)
- [Non renseigné](#) (347)
- [Lille](#) (344)
- [Toulouse](#) (306)
- [missing](#) (38)

**ressource**

- [missing](#) (0)

**Facette d'intervalles**

**valeur**

- [0.0 - 10000.0](#) (7947)
- [10000.0 - 20000.0](#) (1037)
- [20000.0 - 30000.0](#) (434)

**millesime**

10303 résultats trouvés en 24 ms Page 1 de 1031

**Établissement:** Centre national de la recherche scientifique (CNRS)

**Catégorie:** Organisme de recherche

**Catégorie juridique:** Établissement public national à caractère administratif

**Ressource:** RELX - Elsevier - Freedom Collection

**Valeur (en €):** 4563765.0

**Académie:** **Paris**

**Millesime:** 2018

**Département:** Paris

**Région:** Île-de-France

**Wikidata:** Q280413

**Établissement:** Centre national de la recherche scientifique (CNRS)

**Catégorie:** Organisme de recherche

**Catégorie juridique:** Établissement public national à caractère administratif

**Ressource:** RELX - Elsevier - Freedom Collection

**Valeur (en €):** 4430675.0

**Académie:** **Paris**

**Millesime:** 2017

**Département:** Paris

**Région:** Île-de-France

**Wikidata:** Q280413

figure 9 Recherche « Paris » par ordre décroissant

Recherche:

☐ Valeur par ordre décroissant

> [millesime:\[2016 TO 2017\]](#) > [categorie:"Université"](#)

**Facette de champ**

**catégorie**

- [Université](#) (834)
- [missing](#) (0)

**academie**

- [missing](#) (0)

**ressource**

- [missing](#) (0)

**Facette d'intervalles**

**valeur**

- [0.0 - 10000.0](#) (759)

**millesime**

- [2016 - 2017](#) (834)

**Facette de pivot**

- [negociee:false](#) (798)
- [negociee:true](#) (36)
- [Ebst:false](#) (798)
- [Ebst:true](#) (36)

834 résultats trouvés en 6 ms Page 1 de 84

**Établissement:** Université Clermont Auvergne (UCA)

**Catégorie:** Université

**Catégorie juridique:** Établissement public à caractère scientifique, culturel et professionnel

**Ressource:** **Europresse**

**Valeur (en €):** 29690.0

**Académie:** Clermont-Ferrand

**Millesime:** 2016

**Département:** Puy-de-Dôme

**Région:** Auvergne-Rhône-Alpes

**Wikidata:** Q28057967

**Établissement:** Université de Bretagne Occidentale (UBO)

**Catégorie:** Université

**Catégorie juridique:** Établissement public à caractère scientifique, culturel et professionnel

**Ressource:** **Europresse**

**Valeur (en €):** 4051.0

**Académie:** Rennes

**Millesime:** 2016

**Département:** Finistère

**Région:** Bretagne

**Wikidata:** Q1857334

figure 10 Exemple « Europresse »

La figure 10 montre un exemple « chercher Europresse qui en 2016 était la ressource de dépense pour l'université ».

## 9. Conclusion

Solr est une plateforme très utile pour l'indexation et la recherche des informations. Grâce à ce projet, j'ai pu comprendre le principe de fonctionnement de moteur de recherche, comme Google, Bing. Mon moteur de recherche est préliminaire, cependant il est suffisant pour que les utilisateurs puissent le consulter et faire les requêtes. Ce moteur de recherche est visé aux gens qui se focalisent sur l'enseignement supérieur et

sur les ressources électroniques, ou à faire le budget de dépense des ressources électroniques à l'université ou à la bibliothèque.